

The Design of the W. M. Keck Observatory Archive

G. B. Berriman

*Michelson Science Center and Infrared Processing and Analysis Center,
California Institute of Technology, Pasadena, CA 91125, Email:
gbb@ipac.caltech.edu*

D. R. Ciardi

*Michelson Science Center, California Institute of Technology, Pasadena,
CA 91125*

A. C. Laity and N. D. Tahir-Kheli

*Infrared Processing and Analysis Center, California Institute of
Technology, Pasadena, CA 91125*

A. Conrad, J. Mader, and H. Tran

W. K. Keck Observatory, Kamuela, HI 96743

T. Bida

Lowell Observatory, Flagstaff, AZ 86001

Abstract.

The Michelson Science Center(MSC) and the W. M. Keck Observatory are building an archive that will serve data obtained at the Keck Observatory. The archive has begun operations and is ingesting Level 0 (uncalibrated) observations made with the recently upgraded High Resolution Echelle Spectrometer (HIRES); these observations will be publicly accessible after expiration of a proprietary period. Observatory staff have begun using the archived data to determine the long-term performance of the HIRES instrument. The archive is housed at the Michelson Science Center (MSC) and employs a modular design with the following components: (1) Data Evaluation and Preparation: images from the telescope are evaluated and native FITS headers are converted to metadata that will support archiving; (2) Trans Pacific Data Transfer: metadata are sent daily by e-mail and ingested into the archive in a highly fault tolerant fashion, and FITS images are written to DVDs and sent to MSC each week; (3) Science Information System: inherited from the NASA/IPAC Infrared Science Archive, it provides all the functionality needed to support database inquiries and processing of requests; and a Web-based (4) User Interface, a thin layer above the information system that accepts user requests and returns results. The design offers two major cost-saving benefits: it overcomes the geographical separation between the telescope and the archive and enables development at Keck and at MSC to proceed independently; and it permits direct inheritance of the IRSA architecture.

1. Introduction

The W. M. Keck Observatory Archive (KOA) (<http://msc.caltech.edu/koa.html>, and <http://www2.keck.hawaii.edu/realpublic/koa>) is a collaboration between the Michelson Science Center (MSC) and the W. M. Keck Observatory (WMKO). It aims to:

1. Promote the National Aeronautics and Space Administration (NASA) Navigator Program goal of searching for extra-solar planets
2. Curate and disseminate observations made on Keck Single Aperture instruments to maximize science return from the observatory
3. Enable long-term instrument performance studies that will benefit development of observing programs

The KOA entered operations on August 18, 2004, when it began ingesting level 0 (uncalibrated) data obtained with a major upgrade to the High Resolution Echelle Spectrometer (HIRES). The upgrade replaced a single CCD chip with a mosaic of three 2048 x 4096 15 μ m pixel MIT/Lincoln Labs CCDs. The HIRES supports a broad range of astrophysical research, but is most celebrated for its role in the discovery of extra-solar planets.

By November 3, 2004, the KOA had ingested metadata from 54 nights of observations, totaling 6923 observations, and 43 nights (118 GB) of data in Flexible Image Transport System (FITS) format. Currently, access to these datasets is limited to HIRES Principal Investigators (PIs), who are able to query, subset and download their own datasets. The archive contents will be made public after expiration of a proprietary period that is under negotiation between the California Association for Research in Astronomy (CARA) and NASA.

2. Design of the KOA

The science and calibration data, written in FITS format, along with ancillary data, such as summaries of weather conditions and observing logs, are transferred from the observatory to the MSC, where they are curated and served. The KOA employs a modular design consisting of four uncoupled components, as follows:

1. Data Evaluation and Preparation (performed at the Observatory).
2. Trans-Pacific Data Transfer from the Observatory to MSC.
3. Science Information System (maintained at MSC).
4. User Interface (maintained at MSC).

The relationships between these components are shown in Figure 1.

This modular approach to design offers three substantial cost-saving benefits to the archive, as itemized below:

1. It overcomes the geographical separation between telescope and archive.
2. It reduces maintenance costs, as one component can be upgraded independently of the others.
3. It enables inheritance of existing operational software architecture that underpins the NASA/IPAC Infrared Science Archive (IRSA).

The following subsections describe the functions of these components in more detail, identifying where software was reused.

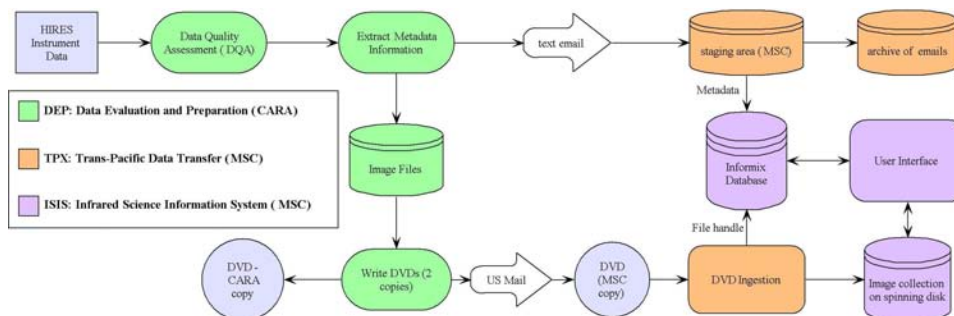


Figure 1. The Design Components of the Keck Observatory Archive.

2.1. Data Evaluation and Preparation (DEP)

This component validates the integrity of the headers and science content of the FITS files created at the telescope. It adds content to the FITS headers to support archiving, as follows:

1. A unique file identifier.
2. An image type identifier that distinguishes program objects from calibration files and identifies the type of calibration file.
3. Data integrity keywords that describe the status of the instrument and its configuration.

Following validation of data, DEP writes all the keywords into a metadata table, with one record for each FITS file (in column delimited format, suitable for ingestion into the database at MSC); places the FITS files and ancillary files in a local disk farm, and copies them to DVDs: one DVD resides at the Observatory and the other resides at MSC.

2.2. Trans-Pacific Data Transfer (TPX)

The metadata are ingested into the archive within 24 hours of observation to support short-term instrument performance analysis by KOA staff at the Observatory. The metadata are transferred via text email to a staging area at MSC. Following automated validation of the content of the metadata, they are ingested into an Informix database within 24 hours of observation, and the emails themselves are then archived.

The FITS files and ancillary data are ingested within 30 days of the observations. The data are written on DVDs that are sent to MSC via U.S. mail, where the contents are copied to a staging area and validated in two ways: by comparison of the MD5 checksums with values of checksums included on the DVDs, and by comparison of the datatypes and values of the metadata in the FITS files with those in the database. The files are then copied to the configured archive disk farm, and file handles are attached to the metadata entries in the database.

The TPX processes are essentially wrappers that automate ingestion tools developed by IRSA that have up to now been used for ingestion of static datasets. The content of the FITS keywords and the structure of the data on the DVDs are managed through an interface control document.

2.3. Science Information System and User Interface (UI)

The KOA has inherited the component-based architecture used that underpins the NASA/IPAC Infrared Science Archive (IRSA¹), which has been operational since spring of 1999. The architecture has been designed to archive and serve the types of data used by astronomers, such as FITS files and ASCII tables. Its applicability is unrestricted by wavelength, despite being developed specifically for an infrared archive, because it consists of a collection of tools or libraries, each of which performs a specific task. The architecture is optimized for two-dimensional astronomical spatial searches by applying a Hierarchical Triangular Mesh² as a spatial indexing scheme.

User services are developed by linking existing modules through a simple executive program. New functionality is added as required. While invariably driven by the needs of a specific data provider, this functionality is written for generality, to maximize the flexibility of the archive services. There are two such augmentations made to support the KOA:

1. *Secure PI Access.* The KOA leverages the Request Object Management Environment, ROME (Good, Kong and Berriman 2004), built at IPAC to support the National Virtual Observatory. ROME is middleware designed expressly to manage user access and time intensive jobs, but the KOA takes advantage only of the user access capabilities. ROME is written as Enterprise Java Bean servlets running under the JBoss Open Source Application Server. It uses Java Data Base Connectivity (JDBC) to provide an interface to an Informix database, which contains a permanent store of authorized users, their passwords and a handle to a permanent workspace that acts as a staging area for retrieved data.
2. *Calibration File Association* The KOA has implemented simple algorithms for associating calibration files with science files. Once science files satisfying a query have been retrieved, the algorithms find the calibration files that have the same instrument and CCD configuration parameters, within specified tolerances.

3. KOA Services to Users

3.1. Secure Access for Keck Principal Investigators

In a given semester, KOA assigns and stores each PI with a user name and password that is stored by ROME and passed on to PIs after the first night of a run. PIs are automatically notified when their data have been ingested into the archive.

3.2. User Interface

Through a simple web form, the KOA supports queries for science and/or calibration files by target name, position (in all common coordinate systems) and radius, date and time, observation parameters (e.g., exposure time, wavelength

¹<http://irsa.ipac.caltech.edu>

²<http://www.sdss.jhu.edu/htm>

coverage), and program information (e.g., PI, program title). The current interface is an advanced prototype that is expected to evolve according to usage patterns and comments from PIs.

The interface offers three options: return science files, calibration files, or the science files and associated calibration files that share instrument and CCD configuration parameters with the science files, as described above. Ancillary data are not yet accessible through the user interface.

The KOA also supports a Structured Query Language (SQL) interface that allows completely general queries of all metadata fields that are stored in the database. Currently, this interface is accessible only by MSC and WMKO archive staff, who are exploiting it to begin investigations of the long term performance of the HIRES instrument.

3.3. Query Results Page and File Download

The web interface returns an HTML page that lists in separate tables all science and calibration files that satisfy the input query criteria. They are available for download one at a time, or through a packaging mechanism that stages a tarball of selected files on a URL-accessible staging area and informs the user by email that the file is ready for download. The package contains, as applicable, science and calibration files, and tabulations of associated calibration files.

4. Planned Functionality

The archive plans to implement the following functionality:

1. More efficient calibration grouping algorithms that overcome the I/O limitations of the current implementation
2. Queries at multiple input positions
3. Access to ancillary data
4. Access to quick-look products
5. Compliance with National Virtual Observatory Standards
6. Support for data tagging initiatives currently being developed by NASA data centers and astronomical journals

Acknowledgments. The W. M. Keck Observatory Archive is funded by the Navigator Program of the National Aeronautics and Space Administration (NASA).

References

- Kong, M., Good, J. C. and Berriman, G. B., 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 213